Procesamiento del Lenguaje Natural, Revista nº 65, septiembre de 2020, pp. 67-74 recibido 03-04-2020 revisado 26-05-2020 aceptado 09-06-2020

Edición de un corpus digital de inventarios de bienes

Edition of a digital corpus of inventories of goods

Pilar Arrabal Rodríguez
Universidad de Granada
pilararrabal@ugr.es

Resumen: En este trabajo se pretende dar a conocer el proceso de elaboración de un corpus diacrónico digital a partir de la selección y edición digital de inventarios de bienes de los siglos XVIII y XIX de las provincias de Madrid y Almería. Los recuentos de bienes, de estructura repetitiva y abundantes en distintos puntos de la geografía hispánica, facilitan la comparación regional y cronológica de los documentos. Este corpus forma a su vez parte de *Oralia diacrónica del español* (ODE), un corpus que toma como inspiración el modelo tecnológico empleado por el proyecto europeo *P.S. Post Scriptum* para ofrecer en línea un corpus anotado a partir de la herramienta TEITOK (Janssen, 2016).
Palabras clave: Inventario de bienes, Lingüística histórica, Lingüística de corpus, español, XML, TEITOK

Abstract: The aim of this paper is to show the process of preparing a digital diachronic corpus based on the selection and digital edition of inventories of goods from the 18th and 19th centuries in the provinces of Madrid and Almeria. Inventories of goods which have a similar structure and are abundant in various areas of the Spanish geography, facilitate the comparison of the documents. This corpus is part of *Oralia diacrónica del español* (ODE), a corpus that takes as inspiration the technological model used by the European project *P. S. Post Scriptum* to offer an annotated corpus online using the TEITOK tool (Janssen, 2016).
Keywords: Inventories of goods, Diachronic linguistics, Corpus Linguistics, Spanish, XML, TEITOK

1 Introducción

En este trabajo se explica la metodología que se está siguiendo para elaborar un corpus diacrónico digital constituido por inventarios de bienes de los siglos XVIII y XIX de las provincias de Madrid y Almería.

Son cada vez más numerosos los corpus basados en documentación archivística que han sabido aprovechar las ventajas que ofrecen este tipo de documentos. En lengua española merece especial mención el corpus CHARTA[1], que integra numerosos subcorpus de muy diferentes tipologías, desde los albores del castellano hasta la época moderna. Entre otras ventajas, la documentación archivística posibilita con sobrada fiabilidad el estudio de la variación diatópica y diacrónica, lo que permite establecer diferencias geográficas claras en el uso de las palabras, así como delimitar la zona y extensión de los términos a lo largo del tiempo. De manera particular, los inventarios de bienes se benefician también de estas condiciones favorables para el estudio de fenómenos dialectales reflejados por los escribanos locales.

La gran abundancia de este tipo de documentos por todo el territorio hispánico permite comparar los resultados tanto diatópica como cronológicamente en la medida en que es posible ubicar con certeza el uso de los vocablos en un momento histórico y lugar geográfico concretos. Un claro ejemplo de ello es el *Corpus Léxico de Inventarios*[2], (Morala

[1] http://www.charta.es/

[2] http://web.frl.es/CORLEXIN.html

ISSN 1135-5948. DOI 10.26342/2020-65-8

Rodríguez, 2014). El *CorLexIn* reúne inventarios de bienes del siglo XVII de prácticamente casi todas las provincias de España y de otras partes de América. De esta manera, queda registrada de forma minuciosa una gran cantidad de léxico referido a objetos de muy diversa naturaleza que eran de extendido uso en la época y en todas las regiones hispánicas favoreciendo así interesantes investigaciones[3].

La abundante documentación conservada y la similar estructura que mantienen los textos, al igual que su contenido, e indistintamente de cuál sea su origen, favorece la investigación dialectal. Estas características hacen de los inventarios una interesante fuente documental que favorece la comparación.

Dejando a un lado los aspectos lingüísticos y pasando a las consideraciones tecnológicas, la edición digital de los documentos que conforman el corpus ha seguido de cerca el modelo tecnológico propuesto por el proyecto europeo *P.S. Post Scriptum*[4]. Este corpus reúne una amplia colección de cartas privadas tanto en español como portugués de autores de diferentes categorías sociales. Con ello se prioriza el estudio de variedades lingüísticas que no suelen aparecer reflejadas en otros corpus de carácter culto o literario, ya que la correspondencia privada se presta a evidenciar un discurso cercano a la oralidad por el tratamiento de asuntos cotidianos (Vaamonde, 2015).

Adicionalmente, el principal objetivo de este proyecto ha sido el de ofrecer un corpus enteramente editado, lematizado y anotado lingüísticamente gracias a la herramienta TEITOK (Janssen, 2016).

En la actualidad ya son veintidós los proyectos que utilizan TEITOK para alojar sus corpus. De entre ellos, *P. S. Post Scriptum* es el corpus que se ha tomado como referencia para la realización del nuevo corpus de inventarios por tratarse también de un corpus histórico de la Edad Moderna del español.

Además del empleo de TEITOK, el uso de estándares internacionales es imprescindible para no dar cabida a textos en los que la presentación gráfica sea heterogénea debido a diferentes criterios que además son desconocidos por el usuario que consulta el corpus. Para la elaboración de este corpus se ha tomado el estándar propuesto por el consorcio TEI para la transcripción de manuscritos.

El corpus que aquí se presenta justifica su elaboración en el principio de la fiabilidad de los documentos que lo componen y pretende ser una herramienta interdisciplinar que facilita investigaciones en diversas áreas como pueden ser la etnografía, historia, cultura y lingüística.

En definitiva, la idea que ha motivado la elaboración de un corpus como el que se propone parte de dos conceptos que se consideran clave: el aprovechamiento de la tipología inventario de bienes tal y como se consolida en *CorLexIn*, y el empleo de un modelo tecnológico para ofrecer un corpus con los últimos avances como lo es el propuesto por *P. S. Post Scriptum*.

2 Antecedentes

El origen del corpus cuya realización se expondrá en los apartados siguientes supone la continuación del proyecto *Corpus diacrónico del español del Reino de Granada. 1492-1833, CORDEREGRA* (Calderón-Campos y García-Godoy 2010), que recoge, entre otras tipologías de documentos, inventarios de bienes de las provincias de Granada, Málaga y Almería. No obstante, el número de textos era considerablemente menor en el caso de esta última. Por tanto, el objetivo ha sido completar el volumen del corpus con la inclusión de documentos de Almería para mostrar así un corpus completo, más rico y representativo de esta región.

Como mejoras frente al corpus ya existente, la ampliación prevé el uso de un avanzado modelo tecnológico que permitirá ofrecer un corpus enteramente digital gracias a la herramienta TEITOK. El resultado tras este cambio en la metodología es el corpus *Oralia diacrónica del español* (ODE) (Calderón Campos, 2019). El corpus que aquí se presenta es por tanto un subcorpus dentro de ODE.

La aplicación de una de las tecnologías más punteras en lingüística de corpus es el objetivo principal de los proyectos en los que este trabajo se enmarca: *HISPATESD* y *ALEA-XVIII*, que pretenden además alcanzar un elevado impacto social con la visibilidad y recuperación del patrimonio documental de Andalucía (por ahora de sus provincias

[3] Morala Rodríguez (2014) ha evidenciado en numerosos trabajos cómo corpus de estas características pueden hacer útiles aportaciones al estudio del léxico y a la lexicografía histórica.

[4] http://ps.clul.ul.pt/es/index.php?

orientales y más adelante de su totalidad). El segundo proyecto, como una de sus aplicaciones tecnológicas más relevantes, persigue además documentar formas léxicas identitarias de todas las provincias andaluzas a partir de un cartografiado diacrónico.

Se pretende contrastar la modalidad lingüística de la región con otras zonas y épocas, por lo que se ha proyectado la creación de un corpus de control con el que poder establecer diferencias o similitudes comparativas. Este corpus será representativo de la Comunidad de Madrid y servirá como referencia en tanto que reflejará una modalidad de habla lo suficientemente alejada de la registrada en Almería. De aquí en adelante y salvo que se especifique lo contrario, se hará referencia de manera conjunta a la metodología empleada tanto en el corpus de estudio almeriense como en el madrileño.

3 El corpus documental

La primera tarea para la confección del corpus se fundamentó en localizar, seleccionar y digitalizar los manuscritos. Para ello se consultaron los fondos notariales del Archivo Histórico de Protocolos de Almería y de Madrid. La labor de selección de los inventarios de bienes ocupó los primeros meses desde el inicio de elaboración del corpus y actualmente está concluida para ambas provincias.

En el caso de los documentos almerienses seleccionados, estos proceden de los once municipios ya contemplados en *CorLexIn* con el fin de mantener lo más fielmente posible el principio de la comparabilidad diatópica. A estos once puntos se le han añadido ocho nuevas localizaciones extendiendo así el espacio estudiado.

A su vez, el corpus de Madrid incluye inventarios de la capital y de otros trece municipios de la provincia. De esta manera se consigue un corpus dialectalmente más amplio y representativo de la provincia en su totalidad de lo que podemos encontrar en *CorLexIn*[5].

La selección de los inventarios de bienes no ha seguido ningún tipo de restricción a excepción de que estos se incluyan en el marco cronológico (XVIII-XIX) y la zona geográfica (Madrid y Almería) objetos de estudio. Ante la abundancia de este tipo de documentación en los archivos, se han preferido los manuscritos en mejor estado de conservación y solo se han rechazado aquellos inventarios que han planteado alguna duda sobre su certera datación o localización.

Los documentos seleccionados son, en su gran mayoría, cartas de dote y recuentos de bienes realizados con motivo de la muerte de algún familiar y tras los que se realiza su consecuente reparto entre los herederos; pero también se incluyen entre los documentos escogidos almonedas o embargos de bienes. En todos ellos el escribano recoge listas pormenorizadas de las pertenencias del benefactor, entre los que se encuentran utensilios domésticos variados, ropas, mobiliario o animales. En la Tabla 1 se muestra una tipología de los documentos pertenecientes a cada provincia incluidos en el corpus:

	Almería	Madrid
Cartas de dote	5	26
Particiones	44	18
Almonedas	2	1
Embargos	1	0
Total	52	45

Tabla 1: Relación de documentos por provincia

A la tarea de selección, le ha seguido la reproducción fotográfica del documento. Cada inventario se ha identificado con una cabecera en la que se incluyen datos que registran su datación y localización geográfica entre otras características descriptivas propias del documento.

Un propósito inicial radicaba en alcanzar un corpus de aproximadamente 100.000 palabras transcritas. Actualmente, con los cerca de cien documentos que componen el corpus, el material digitalizado sobrepasa este tamaño por lo que no se descarta seguir ampliando las dimensiones del corpus en fases posteriores.

4 Transcripción en XML-TEI

Los criterios de transcripción escogidos han sido el resultado de un largo proceso de reflexión y de toma de decisiones con el fin de lograr una edición digital que asegure el rigor filológico de las transcripciones. Tan solo se han normalizado la puntuación, el uso de mayúsculas y minúsculas y la separación de palabras, todo ello conforme a las normas ortográficas actuales. Las grafías se han

[5] La Comunidad de Madrid en *CorLexIn* únicamente se ve representada por tres de sus localidades, entre las que se incluye la capital.

respetado conforme al original con el fin de permitir estudios de carácter gráfico o fonético.

En cuanto a cuestiones de carácter técnico, la edición del total de documentos que componen el corpus se ha llevado a cabo en un lenguaje informático y versátil como es XML adaptado al estándar TEI (Text Encoding Initiative, 2007), un esquema de codificación muy especializado y de alcance internacional.

A partir de este tipo de lenguaje, el consorcio TEI propone un estándar que ofrece una metodología de codificación para la edición de textos en el ámbito de las Humanidades, y más particularmente para la marcación de manuscritos, que es la que ha servido de base para la edición de este corpus.

La transcripción en XML permite combinar aspectos de contenido y formato del documento gracias al etiquetado de diferente información, tal y como se muestra en la Figura 1:

```
        <pb n="3v" facs="20180601_132858.jpg"/>
        <p><lb/> <add place="margin">Imbentario</add> Estando en
mortuoria de Pedro Sevilla Cortinas, <lb/> vecino q<ex>u</ex>e fue
de las Cuebas, sita en la calle de los Na<lb break="no"/>bos de su
el d<ex>ic</ex>ho día veinte de sep<ex>tiembr</ex>e de mil <lb/> s
ochenta y seis, d<ex>o</ex>n Luis Flores Navarro, alg<ex>uaci</ex>
m<ex>ayo</ex>r <unclear>d este</unclear> juzgado, en uso de la com
q<ex>u</ex>e le está con<lb break="no"/>ferida, con asistencia de
es<ex>criba</ex>no y de Indalecio de <lb/> Meca y Josef Sevilla Ga
interesadas, <lb/> procedió a la práctica del imbentario mandado d
vienes q<ex>u</ex>e se manifiestan, a dejado el referido difun<lb
en la forma siguiente:
            <lb/> Prim<ex>eramen</ex>te una colcha de lana, azul,
encarnada.
            <lb/> It<ex>em</ex>: siete sávanas de lienzo almarieta
            <lb/> It<ex>em</ex>: tres pares de calzoncillos del mi
            <lb/> It<ex>em</ex>: cuatro camisas para hombre.
            <lb/> It<ex>em</ex>: otro par de calzoncillos.
            <lb/> It<ex>em</ex>: dos chamarretas blancas.
```

Figura 1: Transcripción en XML-TEI de un fragmento de inventario

En el fragmento transcrito se visualiza cómo es posible marcar características referidas a la estructura principal del documento con las etiquetas <pb/>, <p> o <lb/>; correspondientes a los inicios de folio, párrafo o línea respectivamente. También características físicas o visuales como la que refleja la etiqueta <add> para las adiciones en el margen fuera de la caja de escritura; o bien características conceptuales por medio de las etiquetas <ex> y <unclear>, para el desarrollo de abreviaturas o conjeturas editoriales respectivamente.

La edición en XML se ha llevado a cabo utilizando el procesador de textos *oXygen XML Editor*[6] que incorpora multitud de plantillas para trabajar con los distintos esquemas existentes de codificación, entre los que se encuentran las directrices propuestas por el consorcio TEI. El uso de *Oxygen* ha facilitado sobremanera el etiquetado y ha agilizado el proceso de transcripción, ya que permite validar automáticamente los documentos de acuerdo con el esquema elegido y asiste al usuario en la codificación.

La transcripción de los inventarios se ha centrado exclusivamente en aquellas partes consideradas de mayor interés léxico. Estas son las coincidentes con las listas o recuentos de bienes donde se enumeran pormenorizadamente los objetos de los que consta el inventario. Se han excluido, por tanto, aquellas partes referidas a las relaciones de parentesco de los herederos con el difunto o la relación de parientes implicados en la tasación y reparto de bienes. En definitiva, se han obviado fragmentos donde abundan las fórmulas fraseológicas o los tratamientos protocolarios propios de textos de carácter notarial como son los documentos que nos ocupan. Estas partes resultan claves para la correcta identificación del inventario y por ello se han digitalizado y conservado internamente, pero no se suman al conjunto del material transcrito. De manera excepcional, sí se han transcrito cuando brevemente permiten la identificación del documento y aparecen inmediatamente antes de la relación de bienes.

5 Procesamiento lingüístico del corpus

La edición digital del conjunto de datos no es la meta final. El tratamiento lingüístico del corpus supone uno de los grandes objetivos de carácter tecnológico con el que poder ofrecer un corpus anotado. El procesamiento del corpus consta principalmente de cuatro tareas que tienen que ver con la tokenización, normalización, lematización y etiquetado morfosintáctico. Todas estas tareas se realizan en línea directamente desde el sistema web TEITOK.

TEITOK está ideado para combinar la edición filológica con los avances de la lingüística computacional y abarca dos recursos en uno. En primer lugar, es una plataforma de consulta en la que cualquier usuario externo puede visualizar los documentos y realizar búsquedas en el corpus ajustadas a sus necesidades. En segundo lugar, es también donde se llevan a cabo por parte del equipo de trabajo las tareas de tratamiento lingüístico y

[6] https://www.oxygenxml.com/

edición del corpus que se desarrollan a continuación.

5.1 Tokenización

La tokenización supone uno de los primeros pasos para el tratamiento automático de los textos y alude al proceso de identificación de tokens. Esta tarea se lleva a cabo automáticamente en TEITOK una vez que los documentos han sido importados a la plataforma. El proceso de tokenización consiste en asignar a cada forma ortográfica, incluyendo también los signos de puntuación, un número único de identificación dentro de un elemento <tok> </tok> que engloba la palabra y cualquier otra información que se añada en relación con ella. A partir de ahora, dentro de cada token se almacenará toda la información relativa a los distintos niveles de edición de una sola palabra a la que posteriormente se le adjudicarán los atributos @form (correspondiente a la forma original), @fform (forma expandida) y @nform (forma normalizada), de manera que se respetan las múltiples formas gráficas que tiene una misma palabra. Véase el ejemplo de la Figura 2 tomado del token "vezino":

```
<tok id= "w-428" form= "vezno" fform= "vezino"
    nform= "vecino" >vezno</tok>
```

Figura 2: Ejemplo de token

La transformación de palabras en tokens es también imprescindible para los procesos de búsquedas del corpus con el sistema CQP (Corpus Query Processor). Gracias a la tokenización quedan vinculadas todas las formas de una misma palabra en sus distintos niveles de edición. En la búsqueda, el usuario puede decidir qué forma ortográfica desea buscar (paleográfica o normalizada) y el sistema recuperará todas las soluciones que coinciden en su normalización independientemente de las múltiples variedades gráficas que estén presentes en el texto original. Esto permite recuperar todas las variantes formales, incluyendo aquellas poco predecibles, a partir de una sola búsqueda.

5.2 Normalización

La normalización del corpus abarca exclusivamente el nivel ortográfico del texto. La gran variedad ortográfica que presentan los manuscritos obstaculiza sobremanera la posibilidad de ofrecer un corpus anotado morfosintácticamente. La normalización ortográfica supera esta barrera, pero, además, ello hace accesible el corpus al público interesado que no posea conocimientos lingüísticos específicos.

En esta segunda fase se añade la normalización de aquellas grafías que previamente se respetaron en la edición de acuerdo con el manuscrito original por poseer interés filológico y lingüístico. En TEITOK es posible realizar automáticamente la modernización ortográfica del corpus de acuerdo con parámetros previos importados a la plataforma y que sirven de entrenamiento. Tras la normalización automática se añade a cada token un atributo @nform tal y como se ha visto en el ejemplo de la Figura 2.

Tras la normalización automática se realiza una revisión manual conforme a las normas ortográficas actuales de aquellas palabras que no han sido procesadas correctamente. Esta revisión es de vital importancia, pues evita que errores en este nivel se sucedan en las siguientes fases de anotación.

5.3 Anotación lingüística

La anotación del corpus se lleva a cabo a partir de la forma normalizada de cada palabra. El proceso de anotación automática consiste en la adición de dos nuevos atributos @pos y @lemma a cada token ya formado. A estos atributos se incorpora la etiqueta morfosintáctica y el lema correspondiente.

El etiquetario utilizado para anotar el corpus se ha basado en el ya manejado por el proyecto *P. S. Post Scriptum*, aunque con ligeras modificaciones que simplifican en algunos casos las etiquetas utilizadas. Este, a su vez, sigue las directrices del estándar propuesto por el grupo EAGLES para la anotación de lexicones en lengua española.

La anotación morfosintáctica en TEITOK se lleva a cabo con el anotador automático NeoTag (Janssen, 2012), un analizador probabilístico del tipo HMM que adjudica a cada palabra la etiqueta correspondiente según la función gramatical que cumple en el texto. NeoTag funciona calculando la probabilidad de la etiqueta lingüística POS (part-of-speech) para cada forma ortográfica teniendo como base un corpus de entrenamiento. NeoTag busca en él la

frecuencia de aparición de cada palabra con su respectiva etiqueta POS y en base a ella adjudicará una etiqueta u otra dependiendo de la probabilidad.

En aquellas palabras nuevas o desconocidas para las que no hay datos en el corpus de entrenamiento, el analizador escogerá la etiqueta lingüística según otros factores, como son la terminación de la palabra o contextos similares de aparición. Sin embargo, hay otros casos también propensos a errar en la anotación. Se trata de aquellas palabras que pueden presentar ambigüedad léxica: palabras que, aun estando presentes en el corpus, requieran una anotación morfosintáctica distinta a la registrada por hallarse en un contexto diferente.

Hasta alcanzar las 100.000 palabras, el corpus de entrenamiento que sirvió como base para la anotación ha sido el del proyecto *P. S. Post Scriptum,* constituido por cartas de la Edad Moderna. No obstante, dadas las características individuales del corpus epistolar frente al de inventarios, existía un alto porcentaje de error.

Los siguientes ejemplos reales tomados del corpus servirán para ejemplificar cómo funciona el analizador. En la oración "un perol de cobre", que se encuentra fácilmente en el corpus, *cobre* es etiquetado reiteradamente con la etiqueta VMSP1S0 y con el lema *cobrar*. Se trataría por tanto de un verbo. Esto se debe a que en el corpus de entrenamiento (heredado de *P. S. Post Scriptum*) es alta la frecuencia de *cobre* como una forma verbal y no como el elemento químico al que se hace referencia en nuestro corpus.

De manera paralela a la anotación morfosintáctica se lleva a cabo la lematización, también de forma automática con Neo Tag. Para las palabras desconocidas, delimitar la terminación de las que no lo son es la estrategia usada para asignar el lema. Una vez que el analizador detecta varias posibilidades, elegirá la más frecuente. Este caso se aplica a la palabra "céntimos", anotada automáticamente como un verbo y con el lema **centimar*. Esta palabra no se encuentra en los parámetros y se anota conforme a su terminación correspondiente a la flexión verbal de otras palabras que en cambio sí localiza en los parámetros. En muchos casos como este una mala lematización está vinculada a una etiqueta POS errónea, pero en otros, como el caso de *veces*, (correctamente anotado como sustantivo, pero lematizado como *vec), se debe únicamente al procedimiento de asignación del lema. El algoritmo para ello genera en el acto un patrón de análisis morfológico con el que modificar la forma original para obtener el lema (Janssen, 2012). Este sistema fallará por ejemplo en aquellos casos de verbos con flexión irregular o de formas como la que se acaba de mencionar[7].

Tal y como se deduce a partir de estos ejemplos, los casos con mayor probabilidad de error radican en aquellas palabras que pueden presentar diferentes categorías gramaticales, casos de desambiguación, lo que exige una revisión manual de las mismas.

El porcentaje de acierto en la anotación dependerá de varios factores, entre los que se encuentran el tamaño y la calidad del corpus de entrenamiento en cuanto a lo que etiquetas correctamente asignadas se refiere. Pero sin duda, la precisión del analizador mejora cuando este utiliza el propio corpus como entrenamiento y no otros parámetros importados. Esto solo será posible cuando el tamaño del corpus lo permita (Janssen, 2016).

Al presente, debido al tamaño ya considerable del corpus, este se ha desligado recientemente de los parámetros de *P. S. Post Scriptum* y el conjunto de inventarios ya incluidos en TEITOK conforman el propio corpus de entrenamiento, en constante actualización según aumenta su tamaño. Con los parámetros de los inventarios como base se ha aumentado considerablemente la tasa de aciertos tras una anotación automática. La cantidad de nuevos datos anotados no es aún suficiente para poder realizar una comparación objetiva respecto a la efectividad de los parámetros previos. Se espera que se reduzcan las labores de revisión manual y que errores como los mencionados más arriba queden solventados, aunque no se han realizado todavía estadísticas concluyentes que valoren los resultados de la anotación en esta nueva fase[8].

Con todo, la tarea de la anotación lingüística demanda todavía una atención considerable y

[7] Aun así, los fallos en los lemas son muy limitados respecto a la anotación morfosintáctica. La lematización ha sido testada con un 95% de acierto en corpus españoles (Janssen, 2012: 3).

[8] Test de eficacia para la evaluación de NeoTag en la anotación morfosintáctica se han realizado sobre varios corpus del español actual con una precisión del 97% (*vid*. Janssen, 2012). En un corpus de los siglos XVIII y XIX como el que nos ocupa, y con un tamaño considerablemente menor, este porcentaje se verá reducido.

exige necesariamente una revisión manual, en algunos casos semiautomática[9]. Hasta el momento, esta revisión se ha llevado a cabo entre dos anotadores del equipo de trabajo y se está elaborando una guía, todavía de uso interno exclusivamente, en la que se contemplen los criterios estipulados con el fin de garantizar la consistencia de la anotación en la totalidad de documentos. De esta revisión depende la calidad del corpus de entrenamiento para la anotación de futuros documentos conforme va creciendo el corpus. Asimismo, también en esta fase radica la posibilidad de realizar búsquedas complejas mediante el sistema CQP que utiliza TEITOK, restringiendo las consultas y ofreciendo también información relativa al lema y a la etiqueta lingüística de palabras en posiciones específicas de la oración, lo que favorece una potente sintaxis de búsqueda.

6 Estado actual y trabajo posterior

Desde inicios de 2019 hasta ahora la tarea de recopilación de los inventarios de bienes está terminada tanto para el corpus de estudio como para el de control. Actualmente, el tamaño del que constan ambos corpus supera las 117.000 palabras conforme a la distribución que se muestra en la Tabla 2:

	Almería	Madrid
S. XVIII	28.826	30.325
S. XIX	25.785	33.017
Total	54.611	63.342

Tabla 2: Número de palabras en el corpus

Ya están disponibles para su consulta pública el volumen de textos anteriormente recogidos en la Tabla 2. Es posible acceder a ellos a través de la dirección web http://corpora.ugr.es/ode/, donde se aloja el corpus en su totalidad. Junto a cada transcripción, se puede consultar la información metatextual que acompaña a cada documento y que conforma la cabecera de cada uno de ellos. Además, está disponible la triple visualización de los documentos. Esto es: la reproducción fotográfica del documento o facsímil digital, la edición paleográfica y la edición normalizada de cada uno de los manuscritos.

Todos los documentos ya se encuentran normalizados y anotados morfosintácticamente. En este momento, las tareas de elaboración del corpus están centradas en la revisión del etiquetado lingüístico con el objetivo de constituir un corpus de entrenamiento consistente y de calidad con el menor número de errores posibles y con el que se vea reducido el actual porcentaje de error del anotador automático.

El corpus de inventarios ya se entrena con sus propios parámetros. De ahora en adelante, una vez desligado de los indicadores de *P.S. Post Scriptum*, se sigue probando la efectividad automática de la normalización y del analizador tras este cambio. TEITOK facilita la edición de la anotación lingüística a partir de búsquedas que permiten aplicar cambios en bloque. Actualmente se está trabajando en una revisión del etiquetado según este método, que, si bien no es del todo automático, agiliza sobremanera el proceso sin la necesidad de hacerlo documento por documento. Se espera que esta tarea se concluya al finalizar el presente año.

Producto de la metodología empleada, el resultado es ofrecer un corpus donde el usuario puede realizar búsquedas cruzadas de los documentos que ya están, además de normalizados, anotados lingüísticamente, y descargar libremente los archivos en formato XML o TXT con la transcripción de cada documento en la versión que desee.

Entre las tareas más inmediatas que se realizarán próximamente se contempla la necesidad de ampliar el tamaño del corpus con nuevo material transcrito una vez que se haya trabajado en la mejora del etiquetado morfosintáctico de los textos anotados. Está prevista también una proyección cartográfica digital que refleje todas las localidades de las que se tienen datos y que conforman el corpus.

Otras tareas que merecen especial atención están relacionadas con las búsquedas en el corpus. Entre ellas, mejorar la interfaz para hacerla intuitiva y accesible a cualquier usuario no especializado. Al mismo tiempo, se proporcionarán otras opciones avanzadas que posibiliten la recuperación de la información por medio de búsquedas complejas que permitan explotar al máximo un corpus del que será posible explorar muy diversos aspectos pertenecientes a diferentes ámbitos de investigación.

[9] Para una descripción de las distintas posibilidades que ofrece TEITOK a la hora de agilizar las correcciones en el etiquetado morfosintáctico consúltese Janssen, Ausensi y Fontana (2017).

7 Conclusiones

En este trabajo se ha dado a conocer el proceso de elaboración de un corpus de inventarios siguiendo los últimos avances en edición digital y lingüística computacional para corpus históricos.

TEITOK es una de las herramientas existentes que posibilita crear corpus a partir de la combinación de las tareas de transcripción y lingüística computacional en una plataforma con diversas funcionalidades donde se permite al mismo tiempo el mantenimiento en línea del corpus. Gracias a la tokenización se conserva toda la información relativa al texto y a su grafía al igual que se incluye información lingüística. Las ediciones que se ofrecen de un mismo texto a partir de una sola transcripción solo son posibles mediante el etiquetado previo de los distintos niveles de edición.

Partiendo de archivos codificados en el estándar TEI para lenguaje XML se ofrece un corpus enteramente digital, de libre acceso y ya disponible en línea.

En otro orden de cosas, el objetivo de un corpus de inventarios de las provincias de Almería y Madrid de las características mencionadas es el de poder realizar estudios lingüístico-estadísticos comparativos y establecer diferencias significativas a partir de los análisis contrastivos resultantes. Por ahora, el tamaño del corpus permite hacer una primera cala en esta dirección, pero el propósito es continuar ampliando el corpus y abarcar otras zonas que continúen enriqueciendo los análisis.

Los inventarios de bienes permiten realizar interesantes investigaciones en términos culturales y lingüísticos, ya que permiten reflejar la modalidad dialectal de la zona. Esto se traduce en un corpus creado para satisfacer no solo los intereses de los filólogos investigadores, sino también los de lingüistas, historiadores, etnógrafos o público interesado sin conocimientos lingüísticos específicos. Este carácter interdisciplinar es posible a la proyección de un corpus con diferentes ediciones de entre las que el usuario puede elegir la que desee y acceder a los textos a partir de un potente motor de búsquedas.

Agradecimientos

Este trabajo se inscribe en el marco de los proyectos de investigación *Hispanae Testium Depositiones HISPATESD*, de referencia FFI2017-83400-P (MINECO / AEI / FEDER, UE) y *Atlas Lingüístico y Etnográfico de Andalucía, s. XVIII. Patrimonio documental y humanidades digitales* (ALEA XVIII) (proyectos I+D+i Junta de Andalucía - FEDER, P18-FR-695).

Bibliografía

Calderón-Campos, M. 2019. La edición de corpus históricos en la plataforma TEITOK. El caso de Oralia diacrónica del español. *Chimera*, 6:21-36.

Calderón-Campos, M. y M. T. García-Godoy. 2010-2019. *Oralia diacrónica del español* (*ODE*). En línea: http://corpora.ugr.es/ode/

CLUL (ed.). 2014. *P. S. Post Scriptum. Archivo Digital de Escritura Cotidiana en Portugal y España en la Edad Moderna*. En línea: <http://ps.clul.ul.pt>

Consorcio TEI, (eds.). 2007. TEI P5: Directrices para la codificación e intercambio electrónico de texto (Versión 1.5.). En línea: <http://www.tei-c.org/Guidelines/P5/>

Janssen, M., J. Ausensi y J. M. Fontana. 2017. Improving POS tagging in Old Spanish using TEITOK. En *Proceedings of the NoDaLiDa 2017 workshop on Processing Historical Language*, páginas 2-6, Gotemburgo, Suecia.

Janssen, M. 2016. TEITOK: Text-Faithful Annotated Corpora. En *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, páginas 4037-4043, Portoroz, Eslovenia.

Janssen, M. 2012. NeoTag: a POS tagger for grammatical neologism detection. En *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, Estambul.

Morala-Rodríguez, J. R. 2014. El CorLexIn, un corpus para el estudio del léxico histórico y dialectal del Siglo de Oro. *Scriptum Digital*, 3:5-28.

Vaamonde, G. 2015. P. S. Post Scriptum: dos corpus diacrónicos de escritura cotidiana. *Procesamiento del Lenguaje Natural*, 55:57-64.